Un sistema para resumen automático de textos en castellano

Pedro Luis Mateo, José Carlos González,
Julio Villena y José Luis Martínez
DAEDALUS, S.A.
Centro de Empresas La Arboleda, Ctra. N-III, Km. 7,300
28031 Madrid
plmateo@iberia.es,{jgonzalez,jvillena,jmartinez}@daedalus.es

Resumen: Este artículo presenta un sistema resumidor para textos en castellano que combina técnicas clásicas dentro del campo del resumen automático con otras menos frecuentes, como son la detección de anáforas y de marcadores discursivos, para paliar la escasa coherencia inherente a este tipo de resúmenes.
Palabras clave: resumen automático, sistema resumidor, extracción de frases, extracción automática, resolución de anáforas.

Abstract: This paper presents a text summarization system for the Spanish language that combines classic techniques in automatic summarization with less frequent ones, like anaphora resolution and cohesive markers detection in order to fight the lack of coherence intrinsic to automatic text excerpts.
Keywords: automatic summarization, summarization system, sentence extraction, automatic extraction, anaphora resolution.

1 Introducción

El propósito de los resúmenes es facilitar y acelerar la identificación de los temas interesantes de entre una gran cantidad de documentos. El objetivo final es salvar un tiempo de lectura necesario para localizar la información requerida en un determinado momento. El problema es que la elaboración de resúmenes consume abundantes recursos humanos. La aplicación de ordenadores a esta tarea viene siendo estudiada desde hace varias décadas; ya en 1958 se empiezan a proponer sistemas de resumen automático (Luhn, 1958) como solución al creciente número de textos técnicos publicados.

Pero es en la actualidad, sobre todo con el crecimiento espectacular de Internet, cuando se ha hecho más perentoria la necesidad de disponer de esta tecnología, lo que ha potenciado su desarrollo.

Este trabajo se enmarca inicialmente en un proyecto universitario de investigación en el campo de la recuperación de información[1], habiéndose completado, probado e integrado en K-Site[2], un producto comercial en el ámbito de la Gestión del Conocimiento. El objetivo, desde el punto de vista investigador, consistía en evaluar diversas estrategias para la elaboración automática de resúmenes de documentos. Para ello, se desarrolló una plataforma software que facilitara el proceso completo de evaluación.

2 Descripción del sistema

Es práctica habitual en los sistemas comerciales de resumen automático aproximar el resumen de un documento mediante su extracto, entendiéndose como tal una colección de las frases más representativas del texto original copiadas sin sufrir apenas modificación. Este enfoque contrasta con el concepto bien

[1] ACORDEON: Aplicaciones Cooperativas de Recuperación de Información, proyecto financiado por el Plan Nacional de I+D, CICYT TEL99-1073-C02.

[2] K-Site ® es una marca registrada por DAEDALUS-Data, Decisions and Language S.A.

conocido de resumen, que supone la redacción de un nuevo texto basado en el original.

El sistema presentado en este artículo realiza también esta aproximación que, pese a parecer simplista, goza de cierta justificación. Según (Kupiec et al., 1995), aproximadamente el 80% de las frases en resúmenes creados por humanos están copiadas tal cual o con pequeñas modificaciones a partir del texto original. Sin embargo, el uso de extractos presenta algunos inconvenientes: una extensión 3 veces mayor que su resumen equivalente (Hovy y Lin, 1997), menor coherencia y mayor probabilidad de redundancia, entre otros.

El sistema consta de cinco módulos, como se ve en la Figura 1: análisis morfosintáctico, ponderación de frases, detección de anáforas, selección de frases y post-procesado del extracto. El proceso de síntesis comienza con el análisis morfosintáctico del documento de entrada. Basándose en esta información y en la presencia de diversas características superficiales, el módulo de ponderación asigna puntuaciones a las frases del texto según su importancia. Además, entrega a su salida una lista de las frases candidatas (un porcentaje dado de las frases más importantes).

El módulo de selección de frases escoge las oraciones candidatas que han obtenido mayores puntuaciones, teniendo en cuenta la longitud deseada del resumen y la presencia de referencias anafóricas. Permite realizar tanto extracción de párrafos completos como de frases sueltas.

Una vez seleccionadas las frases del extracto, el módulo de post-procesado comprueba la presencia de ciertas expresiones o marcadores discursivos al comienzo de las mismas, con el objetivo de editarlas si fuera necesario. A su salida entrega el resumen del documento.

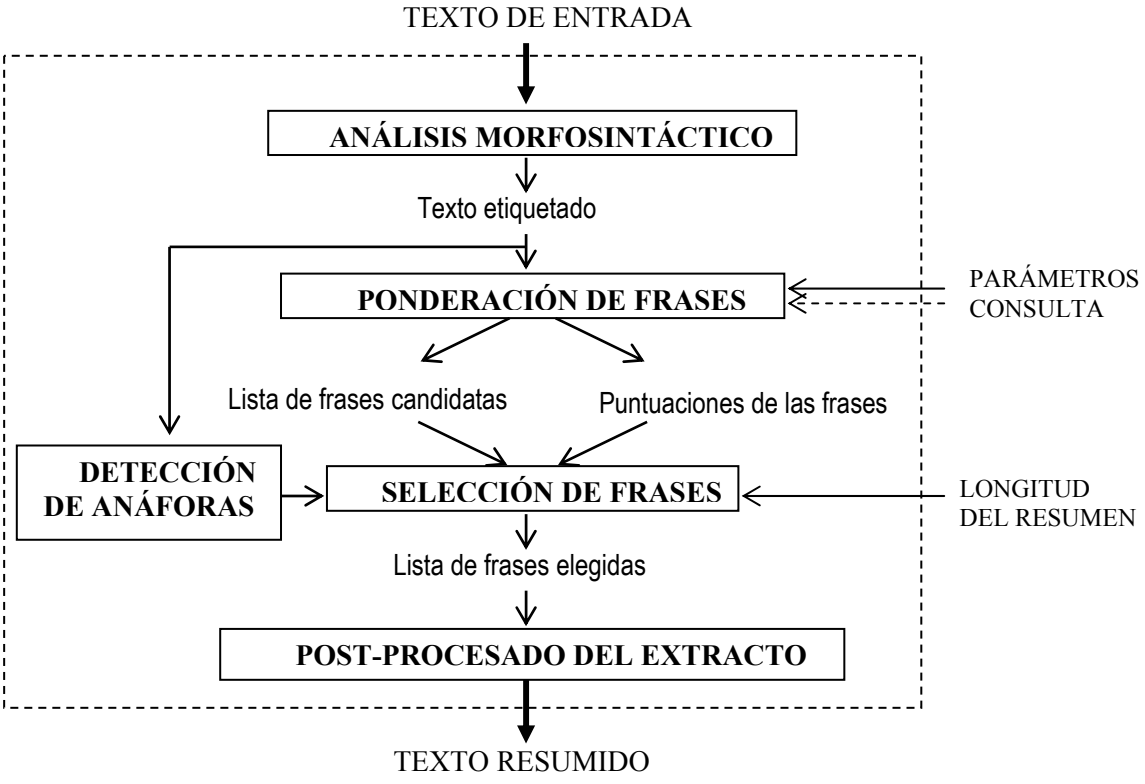

Figura 1: Diagrama de módulos del sistema

2.1 Módulo de análisis morfosintáctico

Este módulo determina la categoría léxica de cada palabra del texto de entrada (sustantivo, artículo, verbo, etc.), así como su lema y formantes. La categoría de las palabras resulta muy útil al módulo de ponderación de frases,

pues permite distinguir entre palabras pertenecientes a clases abiertas (sustantivo, adjetivo, verbo) y cerradas (artículo, pronombre, preposición, etc.). El lema permite considerar como un único concepto todas las formas flexivas de una palabra.

2.2 Módulo de ponderación de frases

Recibe como entrada el texto etiquetado por el módulo de análisis morfosintáctico, los parámetros de configuración del sistema y una consulta de usuario opcional, entregando a su salida una tabla con puntuaciones de las frases y una lista de frases candidatas.

Consta de nueve submódulos, siete de los cuales puntúan las frases basándose en la presencia de ciertos patrones o características superficiales: posición dentro del texto, tipografía, presencia de nombres propios, etc. Los dos módulos restantes son el de combinación de puntuaciones y el de prefiltrado.

2.2.1 Ponderación basada en frecuencias de aparición de palabras y prefiltrado

El principio en que se basa este método es el propuesto originalmente en (Luhn, 1958), si bien el algoritmo es diferente. Sigue la hipótesis de que las palabras que aparecen frecuentemente dentro del documento son relevantes. Cualquier escritor repite ciertas palabras a lo largo del documento, conforme elabora sus argumentos, y esto puede explotarse para calcular el factor de importancia de cada frase.

Se debe tener en cuenta que existen ciertas palabras sin significado cuya función es meramente la de enlazar otras y que, además, suelen aparecer frecuentemente. Son las pertenecientes a categorías léxicas cerradas: determinantes, conjunciones, pronombres, preposiciones, adverbios, verbos auxiliares y verbos modales.

El algoritmo de puntuación implementado comienza con la creación de una tabla que recoge los lemas de todas las palabras del documento pertenecientes únicamente a categorías léxicas abiertas (nombres, adjetivos y verbos léxicos), asignando a cada lema una puntuación proporcional a su frecuencia de aparición. La consideración de los lemas permite agrupar todas las formas flexivas de una palabra en el mismo concepto. De esta tabla inicial se eliminan los de baja frecuencia de aparición con el objetivo de obviar palabras irrelevantes y también errores ortográficos. El umbral utilizado para ello es el número medio de repeticiones de los lemas en el texto.

En la segunda parte del proceso se asigna una puntuación a cada frase. El cálculo se realiza comparando los lemas de las palabras de dicha oración con todos los lemas de la tabla creada anteriormente. Para cada coincidencia se añade a la frase correspondiente la puntuación que tiene ese lema en la tabla y el resultado final se divide por el número total de palabras de la frase, para normalizar respecto a la longitud de la frase. Una vez que se tiene la puntuación de todas las oraciones, se normaliza a uno para facilitar la combinación posterior de las puntuaciones obtenidas por los diferentes métodos.

Este módulo realiza además un prefiltrado de las frases más importantes del documento. Consiste en la eliminación de las frases que han obtenido las puntuaciones más bajas según el método de las frecuencias, considerando como frases relevantes o candidatas un porcentaje dado de las mejor puntuadas.

Adicionalmente, elimina las frases muy cortas porque, de hecho, no suelen incluirse en los resúmenes generados por humanos (Kupiec et al., 1995). Las salidas de este módulo son dos: una tabla con las puntuaciones de cada frase y una lista de frases candidatas.

2.2.2 Ponderación basada en palabras indicativas

Este método pretende encontrar meta-discurso (Teufel y Moens, 1997), es decir, palabras y expresiones que no guardan relación con el tema central del documento, sino que indican que la frase puede contener información importante y debería formar parte del resumen, como por ejemplo: “importante”, “esencial” o “para concluir” (Edmundson, 1969).

Uno de los principales inconvenientes de este método es la dependencia de las palabras indicativas con el género del documento (Hovy y Lin, 1997), por lo que para lograr precisión se necesita una lista de palabras indicativas muy específica del dominio al que pertenece el documento. Dado que se desea un sistema lo más genérico posible, se utiliza un conjunto de

130 palabras determinadas manualmente que revelan importancia independientemente del tipo de documento: “esencial”, “fundamental”, etc.

El algoritmo implementado busca estas palabras dentro del texto a resumir, otorgando puntuaciones a las frases que las contienen.

2.2.3 Ponderación basada en palabras del título

El título de un documento suele estar fuertemente relacionado con su contenido y a menudo constituye el mejor resumen del mismo. Esto es especialmente cierto en artículos periodísticos, aunque es razonablemente válido para otros géneros. Según (Lin y Hovy, 1997), el título puede contener alrededor del 50% de las palabras clave del tema del documento. Por tanto, las palabras del título se pueden utilizar como palabras clave alternativas a las palabras de alta frecuencia.

El algoritmo identifica en primer lugar los lemas de las palabras del título que pertenecen a categorías léxicas abiertas y luego pondera las frases del documento conforme a la aparición en ellas de estos lemas, normalizando por último los valores obtenidos.

2.2.4 Ponderación basada en presencia de nombres propios

Los nombres propios de personas, lugares, organizaciones, etc., proporcionan usualmente información y su inclusión en el resumen puede aumentar la cantidad de datos facilitados al lector. Según un estudio realizado por Goldstein et al. (1999), la densidad media de nombres propios en las frases pertenecientes a resúmenes a menudo es mayor que en las frases no pertenecientes a resúmenes.

El algoritmo puntúa a las frases del documento que contienen nombres propios, normalizando a continuación los valores obtenidos.

2.2.5 Ponderación basada en la tipografía del texto

Es habitual utilizar la tipografía y el formato de un texto para conceder deliberadamente más importancia a unas frases que a otras. Por ejemplo, empleando un tipo de letra diferente, un tamaño de letra más grande, o marcando en negrita el texto se puede destacar ciertas partes del documento.

En concreto, el sistema comprueba la presencia de palabras en mayúsculas, negrita y subrayadas, puntuando a las frases que las contienen.

2.2.6 Ponderación basada en la posición dentro del documento

Este método explota el hecho de que en algunos géneros, las regularidades de la estructura del discurso o la propia forma de exponer la información hace que ciertas posiciones tiendan a contener puntos importantes del documento. Los párrafos iniciales y finales de un documento suelen contener gran riqueza semántica, siendo menos importantes los párrafos centrales (Teufel y Moens, 1997). A su vez, los párrafos también están organizados jerárquicamente: la frase inicial y la final habitualmente son más importantes que las frases centrales.

El algoritmo puntúa las frases conforme a una función lineal decreciente de pendiente configurable que asigna puntuación máxima a la primera frase.

2.2.7 Ponderación basada en una consulta de usuario

El sistema puede obtener resúmenes personalizados sobre un tema concreto, al permitir usar una consulta opcional definida como una serie de “palabras clave” que introduce el usuario y sirve para guiar el proceso de síntesis. Este método concede mayor importancia a las frases que contienen estas palabras, teniendo en cuenta además sus lemas.

2.2.8 Módulo de combinación de puntuaciones

El uso independiente de los heurísticos implementados por los siete métodos anteriores funciona adecuadamente con textos pertenecientes a dominios particulares. Para lograr mayor robustez y generalidad conviene combinar los resultados de estos métodos. Cada uno de los siete módulos de ponderación presentados proporciona una serie de puntuaciones normalizadas a 1 para cada frase

del documento. El módulo de combinación de puntuaciones agrupa todos estos valores en uno único para cada frase, utilizando para ello una función de combinación lineal que suma para cada frase las puntuaciones de los siete métodos ponderando cada uno de ellos por un peso determinado empíricamente mediante pruebas.

2.3 Módulo de detección de anáforas

Al realizar un resumen concatenando frases sin más, el texto resultante no tiene por qué ser necesariamente coherente. Una de las maneras en que se manifiesta esta falta de coherencia es en términos de anáforas sin resolver: una frase del extracto puede contener una anáfora cuyo referente se encuentra en una frase previa que no ha sido tomada para el resumen.

Este módulo pretende paliar el problema buscando expresiones anafóricas en unos casos concretos, pero muy importantes: anáforas de demostrativos pronominales y del pronombre personal "ello". Para simplificar, se realiza la suposición de que la aparición de un pronombre en el medio o final de una frase hace referencia a un concepto introducido previamente en la misma frase (referencia intraoracional), mientras que si aparece entre las primeras palabras de la frase se refiere a alguna oración anterior (referencia interoracional). Son de interés especial las referencias interoracionales, pues son las que determinan la necesidad de incluir oraciones adicionales en el resumen.

En concreto, si se encuentra una referencia anafórica dentro de las 6 primeras palabras de una frase, se supone que se refiere a la oración inmediatamente anterior. Corresponde al módulo de selección de frases decidir si debe incluir la frase anterior en el resumen.

2.4 Módulo de selección de frases

Determina qué frases se incluirán en el resumen considerando la información proporcionada por los módulos de ponderación de frases (lista de frases candidatas y puntuaciones de las frases) y de detección de anáforas, teniendo en cuenta si debe realizar extracción de frases o párrafos completos. Considera en primer lugar las frases candidatas, comenzando por la de mayor puntuación. Esta se elige para el resumen si su longitud no sobrepasa el tamaño deseado del extracto, continuando después con las demás frases según puntuaciones decrecientes hasta completar el resumen. Si tras este proceso no se ha completado la longitud del resumen, se continúa con las frases no candidatas.

La información proporcionada por el módulo de detección de anáforas permite seleccionar frases adicionales para el resumen, con la finalidad de aumentar la coherencia. Cada vez que se elige una frase para incluirla en el resumen, se comprueba si contiene expresiones anafóricas. En este caso, se toma la frase inmediatamente anterior siempre que esta sea una frase candidata y no se sobrepase la longitud máxima del resumen.

2.5 Módulo de post-procesado del extracto

Su objetivo es detectar expresiones que usualmente conectan unas partes del texto con otras, como por ejemplo: "por consiguiente" o "sin embargo". La presencia de alguna de estas expresiones al comienzo de una frase indica una relación de dependencia con el texto anterior, ya sea causalidad, contraposición, etc. Si el resumen contiene una frase con una de estas expresiones pero no la frase anterior de la que depende, el resultado será un texto difícil de entender y poco coherente.

El sistema es capaz de identificar más de 750 expresiones al comienzo de las frases del resumen. En caso de encontrarlas en una frase, la acción tomada depende de que la frase inmediatamente anterior forme parte del extracto; en caso afirmativo se deja la frase tal cual, si no, se elimina la expresión. Esta edición se realiza de forma automática por este módulo.

El siguiente ejemplo ilustra el proceso (en negrita aparecen las expresiones encontradas):

> *Un portavoz de RENFE aseguró ayer que la modificación del itinerario de estos trenes era "factible", aprovechando un ramal de servicio que enlaza las líneas de la mitad sur de la red.* ***"De esta manera,*** *los trenes alcanzarían Villaverde por el norte, no por el sur", explicó.* ***Sin embargo,*** *esta fuente aclaró que la compañía no ha*

llegado a estudiar con detalle esta posibilidad.

No es preciso realizar ninguna edición si el módulo de selección de frases ha elegido las tres para formar parte del extracto, o únicamente la primera. Si ha tomado sólo la última, el módulo de post-procesado eliminará la expresión: "Sin embargo". Si se ha tomado la segunda o las dos últimas, es necesario eliminar "De esta manera".

Conviene destacar el potencial de la actuación conjunta de los módulos de detección de anáforas y post-procesado en términos de mejorar la coherencia del resumen. Consideremos un texto que contiene, entre otras, el siguiente par de frases, donde la segunda contiene una expresión (en negrita) y una referencia anafórica (subrayada):

1. La causa inmediata de esta violencia es la campaña del Ejército de Liberación de Kosovo (ELK) para conseguir la independencia de la región.

*2. **Ello** **constituye una de las razones por las que** la OTAN no ha querido "convertirse en la fuerza aérea del ELK".*

Se pueden dar tres alternativas según las oraciones que tome el módulo de selección de frases:

a) Elige ambas frases debido a su elevada puntuación. En este caso no es necesario considerar la presencia de anáforas ni editar la segunda frase porque la primera proporciona el contexto necesario.

b) Elige la frase 1. El resumen sólo consta de esta frase.

c) Por motivos de puntuación debería elegir la frase 2, pero al contener una anáfora conviene tomar también la frase 1 aunque tenga baja puntuación. El criterio seguido es que sólo se añade la frase 1 si es frase candidata. En caso contrario, el módulo de selección de frases descarta la frase 1 y el módulo de post-procesado edita la frase 2 eliminando la expresión en negrita.

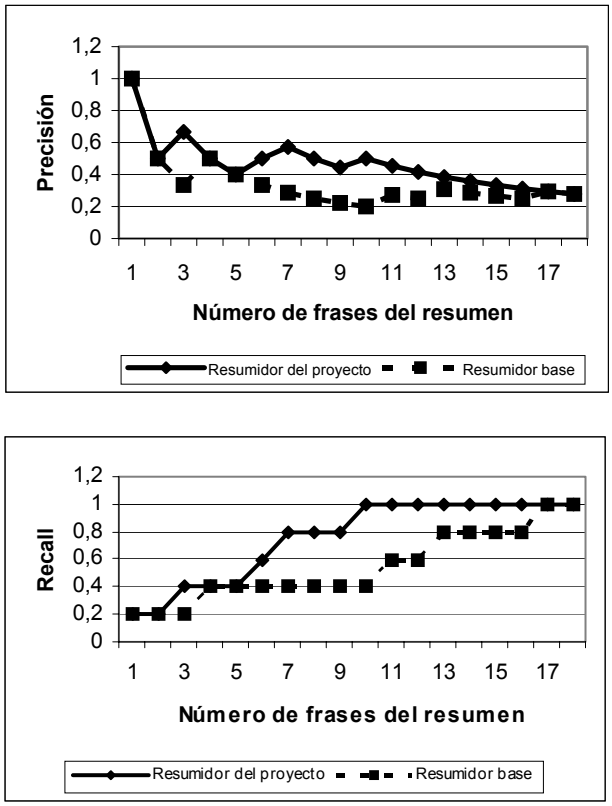

Figura 2: Precisión y *recall* para un texto de prueba

3 Evaluación

Se ha realizado una evaluación estadística basada en extractos "ideales" obtenidos a partir de varios extractos realizados por humanos (en concreto 6 personas). Se han calculado los valores de precisión y *recall* para extractos de diferentes longitudes generados por el sistema. A continuación, se han comparado con los obtenidos por un resumidor base que toma las primeras frases de un documento como su extracto.

En la Figura 2 se ilustran los resultados de precisión y *recall* para un texto concreto utilizado en una de las pruebas, obtenido de un periódico y de una página de longitud (18 frases). En este caso el porcentaje de acuerdo entre los 6 resumidores humanos fue del 81% y la longitud de los extractos manuales cinco frases. Se observa que el sistema expuesto presenta mejor precisión y *recall* que el resumidor base para cualquier longitud de extracto. Estos resultados se mantienen en el

resto de pruebas realizadas con otros textos del corpus de documentos periodísticos utilizado en la evaluación.

4 Conclusiones

Uno de los principales inconvenientes de los sistemas basados en extracción de frases es la escasa coherencia y calidad de los resúmenes que generan. En el diseño del sistema se ha hecho frente a este problema incluyendo dos módulos específicos: detección de anáforas y post-procesado del extracto. Los resultados han sido muy positivos pese a la relativa simplicidad de los métodos.

La resolución de anáforas es un campo complejo que requiere todavía amplia investigación. Permitiría aumentar la precisión de los métodos basados en frecuencias de palabras, así como mejorar la coherencia del resumen, incluyendo la frase que contiene el antecedente de una expresión anafórica. El sistema presentado en este artículo solventa el problema de las anáforas detectándolas en lugar de resolverlas. Las pruebas realizadas demuestran que esto es suficiente en buena parte de las ocasiones.

Otra de las conclusiones obtenidas tras las pruebas consiste en la mayor utilidad del método basado en frecuencias como filtro de las frases más importantes, lo que coincide con los resultados de otros investigadores (Myaeng y Jang, 1999). Parece que la precisión de la tecnología actual es suficiente para distinguir las frases esenciales de las irrelevantes, pero no tan buena como para ordenarlas por importancia relativa.

Una de estas limitaciones de los métodos de extracción de frases empleados actualmente es que, de acuerdo con Zechner (1997), existe un límite superior en las técnicas que consisten de o incluyen palabras. Esto se debe a fenómenos como la sinonimia, polisemia, anáforas, metáforas o metonimia, que generan ambigüedades en el texto que no pueden distinguirse observando palabras sueltas. Una solución es usar bases de conocimiento léxico, como WordNet, para considerar no sólo las frecuencias de palabras sueltas del documento, sino también las de sus sinónimos.

Bibliografía

Edmundson, H.P. 1969. New Methods in Automatic Extracting. *Journal of the Association for Computing Machinery,* 16(2):264-285, 1969.

Goldstein, J., M. Kantrowitz, V. Mittal y J. Carbonell. 1999. Summarizing Text Documents: Sentence Selection and Evaluation Metrics. Carnegie Mellon University.

Hovy, E. y C. Lin. 1997. Automated Text Summarization in SUMMARIST. *ACL Workshop on Intelligent Scalable Text Summarization*.

Kupiec, J., J. Pedersen y F. Chen. 1995. A Trainable Document Summarizer. *Proceedings of the 18th ACM-SIGIR Conference*, páginas 68-73.

Lin, C. y E. Hovy. 1997. Identifying Topics by Position. *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*, páginas 283-290.

Luhn, H.P. 1958. The Automatic Creation of Literature Abstracts. *IBM Journal of Research & Development,* 2(2):159-165, 1958.

Myaeng, S.H. y D.H. Jang. 1999. Development and Evaluation of a Statistically-Based Document Summarization System. *Advances in Automatic Text Summarization.* MIT Press, páginas 61-70.

Teufel, S. y M. Moens. 1997. Sentence extraction as a classification task. *Workshop 'Intelligent and scalable Text summarization'. ACL/EACL*.

Zechner, K. 1997. A literature survey on information extraction and text summarization. Carnegie Mellon University.